The ASA president Task Force Statement on Statistical Significance and Replicability

Yoav Benjamini

Dept. of Statistics & Operations Research
School of Math Sciences and Sagol School for Neuroscience
Tel Aviv University

www. math.tau.ac.il/~ybenja

January 11, 2022

Research supported by NIH-BSF grant
The Reproducibility and Replicability Concern

In Medicine, Soric (‘89)
“There is a danger that a large part of science can thus be made untrue”

In Genomics, Lander and Kruglyak (‘95) warned
“There is danger that a substantial proportion of our claims cannot be replicated”

The beginning of the Industrial revolution in science
The industrialization of the scientific process

1888 1999

1950 2010
Their common argument: select the significant results

180 false effects

20 true effects

\[ \alpha = 0.05 \]

On the average: 9 false discoveries 16 true ones

False Discovery Proportion: \[ \frac{9}{9 + 16} = 0.36 \]
Their common argument: select the significant results

On the average: 9 false discoveries, 16 true ones
False Discovery Proportion: $\frac{36}{36+16} = 0.68$

Landers & Kruglyak: use Bonferroni
$0.05 \rightarrow \frac{0.05}{740}$

Soric: motivated Hochberg and YB work on FDR and BH
The Reproducibility and Replicability Crisis

Psychological Science Journal:
Do not trust any p value. ...
Routinely report 95% CIs...

Basic and Applied Social Psychology (‘15):
*Bans the Null Hypothesis Statistical Testing...*
Is it the p-value’s fault?

American Stat. Assoc. (ASA) Board’s statement about p-values¹ (‘16)

Opens: The p-value “can be useful”

Then comes: a list of “do not” “is not” and “should not”

“leads to distortion” –

All warnings phrased about the p-value

Bayesian vs Frequentists rivalry focusing on p-value,/test/selection
It concludes: “In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. “

It is the p-value’s fault

“We’re finally starting to get rid of the p-value tyranny”
What other approaches were mentioned in ASA statement?

Confidence intervals
Prediction intervals
Estimation
Likelihood ratios
Bayesian methods
Bayes factor
Credibility intervals
Moving to a world beyond ‘p<.05’

Summary of the 43 papers by Wasseerstein, Schirm & Lazar

- Don’t use p<.05
- Don’t say “statistically significant”
- or use any bright-line rule

There are many Do’s

Accept Uncertainty.
Be Thoughtful, Open
And Modest
A Replicability Crisis turned into a Statistical Crisis

The American Statistical Association (ASA) has recently issued two statements on the “p-value < 0.05” criterion for statistical significance widely used in many fields of science: see Wasserstein and Lazar (2016) and Wasserstein et al. (2019). The statements are mainly concerned with the abuse and misuse of the p-value criterion,

Kim, Review of managerial science (2021)
THE ASA PRESIDENT’S TASK FORCE STATEMENT ON STATISTICAL SIGNIFICANCE AND REPLICABILITY

By Yoav Benjamini\textsuperscript{1}, Richard D. De Veaux\textsuperscript{2}, Bradley Efron\textsuperscript{3}, Scott Evans\textsuperscript{4}, Mark Glickman\textsuperscript{5,*}, Barry I. Graubard\textsuperscript{6}, Xuming He\textsuperscript{7}, Xiao-Li Meng\textsuperscript{5,†}, Nancy Reid\textsuperscript{8}, Stephen M. Stigler\textsuperscript{9}, Stephen B. Vardeman\textsuperscript{10}, Christopher K. Wikle\textsuperscript{11}, Tommy Wright\textsuperscript{12}, Linda J. Young\textsuperscript{13} and Karen Kafadar\textsuperscript{14}

This document is the statement of the task force, and the ASA invited us to publicize it. Its purpose is two-fold: to clarify that the use of $P$-values and significance testing, properly applied and interpreted, are important tools that should not be abandoned, and to briefly set out some principles of sound statistical inference that may be useful to the scientific community.
P-values are valid statistical measures that provide convenient conventions for communicating the uncertainty inherent in quantitative results. Indeed, P-values and significance tests are among the most studied and best understood statistical procedures in the statistics literature. They are important tools that have advanced science through their proper application.
The indispensable $p$-value

• It’s the first defense line against being fooled by randomness – needs minimal modeling assumptions

Which can be guaranteed

In a properly designed and executed experiment

• The meaning of $p$-value is shared across fields of science (after adjustment)

• In some emerging branches of science it’s the only way to compare across conditions: GWAS, fMRI, Brain Networks, or make a totally new discovery (Higgs Bosson)
Dealing with replicability and uncertainty lies at the heart of statistical science. Study results are replicable if they can be verified in further studies with new data. Setting aside the possibility of fraud, important sources of replicability problems include poor study design and conduct, insufficient data, lack of attention to model choice without a full appreciation of the implications of that choice, inadequate description of the analytical and computational procedures, and selection of results to report. Selective reporting, even the highlighting of a few persuasive results among those reported, may lead to a distorted view of the evidence. In some settings this problem may be mitigated by adjusting for multiplicity.
Goes back to the 17th century Robert Boyle, during his debate with Huygens over his air pump and the nature of vacuum.
Replicability with significance

“We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results.”

Dealing with replicability and uncertainty lies at the heart of statistical science. Study results are replicable if they can be verified in further studies with new data. Setting aside the possibility of fraud, important sources of replicability problems include poor study design and conduct, insufficient data, lack of attention to model choice without a full appreciation of the implications of that choice, inadequate description of the analytical and computational procedures, and selection of results to report. Selective reporting, even the highlighting of a few persuasive results among those reported, may lead to a distorted view of the evidence. In some settings this problem may be mitigated by adjusting for multiplicity.
Selective inference

Inference on a selected subset of the parameters that turned out to be of interest *after viewing the data*!

Relevant to all statistical methods

1. **Out-of-study selection** - not evident in the published work

   *File drawer problem / publication bias*

   The garden of forking paths, p-hacking, cherry picking

   significance chasing, Data dredging,…

   All are widely discussed and

   Addressed by *Transparent* and *Reproducible* research
Selective inference

In-study selection - evident in the published work:

- Selection by the Abstract, Table, Figure
- Selection by highlighting those passing a threshold $p<.05$, $p<.005$, $p<5\times10^{-8}$, 2 fold
- Confidence / Credible Interval not covering 0 or 1
- Selection by modeling: AIC, $C_p$, BIC, LASSO, ...
- Selection by size of estimated parameter

In complex research problems - in-study selection is unavoidable!
Selection by the Abstract

• Giovannucci et al. (‘95) look for relationships between more than 100 types of food intakes and the risk of prostate cancer

• The abstract reports three (marginal) 95% confidence intervals (CIs), apparently only for those relative risks whose CIs do not cover 1.

“Eat Ketchup and Pizza and avoid Prostate Cancer”

“…this association has remained nonsignificant since 2000 after the addition of 7 studies…”

Rowles et al (‘17)
Schnall et al.’08, presented subjects with 6 moral dilemmas asking “how wrong each action was”

The research goal: Does priming for cleanliness affect the response?

1 Assessment of wrong-doing & 9 emotions rating per dilemma;
Methods of priming verbal & physical (separate experiments) \( (m>84) \)

Results: Only a contrast for disgust, a summary of moral judgment over all dilemmas, and 3 particular dilemmas priming made a significant difference. All tests at 0.05;

Their Conclusion: The findings support the idea that moral judgment is affected by priming for cleanliness.

The replication attempt: failed
The status: Experimental psychology

Analysis of the 100 in the Reproducibility Project in Psychology:

# inferences per study (4-700, average 72); 11 partially adjusted.

56/88 results were not replicable.

After adjusting for selection using hierarchical FDR method

22 should not have been reported as significant ($p_{adj} > 0.05$)

21 non-replicable results appropriately screened

1 replicable discovery was lost

Zeevi, Estachenko, Mudrik, YB (‘21+)
Capturing the uncertainty associated with statistical summaries is critical. Different measures of uncertainty can complement one another; no single measure serves all purposes.

The sources of variation that the summaries address should be described in scientific articles and reports. Where possible, those sources of variation that have not been addressed should also be identified.
“....including a requirement to replace $P$ values with estimates of effects or association and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity. “
“The n−3 fatty acids did not significantly reduce the rate of either the primary cardiovascular outcome or the cancer outcome. If reported as independent findings, the P values for two of the secondary outcomes would have been less than 0.05;
The Abstract of Manson et al 2018

RESULTS A total of 25,871 participants, including 5106 black participants, underwent randomization. During a median follow-up of 5.3 years, a major cardiovascular event occurred in 386 participants in the n-3 group and in 419 in the placebo group (hazard ratio, 0.92; 95% confidence interval [CI], 0.80 to 1.06; P=0.24). Invasive cancer was diagnosed in 820 participants in the n-3 group and in 797 in the placebo group (hazard ratio, 1.03; 95% CI, 0.93 to 1.13; P=0.56). In the analyses of key secondary end points, the hazard ratios were as follows: for the expanded composite endpoint of cardiovascular events, 0.93 (95% CI, 0.82 to 1.04); for total myocardial infarction, 0.72 (95% CI, 0.59 to 0.90); for total stroke, 1.04 (95% CI, 0.83 to 1.31); for death from cardiovascular causes, 0.96 (95% CI, 0.76 to 1.21); and for death from cancer (341 deaths from cancer), 0.97 (95% CI, 0.79 to 1.20). In the analysis of death from any cause (978 deaths overall), the hazard ratio was 1.02 (95% CI, 0.90 to 1.15). No excess risks of bleeding or other serious adverse events were observed.
Wu et al, citing results of Manson et al NEJM 2018
Nature Reviews Cardiology (2019)

Fish oil supplementation ...

had no significant effect on the composite primary end point of CHD, stroke or death from CVD

but reduced the risk of

total CHD* (HR 0.83, 95% CI 0.71–0.97),
percutaneous corona intervention (HR 0.78, 95% CI 0.63–0.95),
total myocardial infarction* (HR 0.72, 95% CI 0.59–0.90),
fatal myocardial infarction (HR 0.50, 95% CI 0.26–0.97).

The only 4 out of 22 that excluded 1. These are no longer 95% CI !
20 parameters to be estimated with 90% CIs

3/20 do not cover

3/4 do not cover when selected

These so selected 4 will tend to fail, or shrink back, when replicated.

FCR CIs have level 

$(1-.1*4/20)100\%$
Are Bayesian intervals immune from selection’s harms?

Assumed Prior $\mu_i \sim N(0,0.5^2); \quad y_i \sim N(\mu_i,1); \quad i=1,2,...,10^6$ (Gelman’s Ex.)

Parameters generated by $N(0,.5^2)$

<table>
<thead>
<tr>
<th>Type of 95% confidence/credence intervals</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervals not covering their parameter</td>
<td>5.0%</td>
</tr>
<tr>
<td>Intervals not covering 0: Selected</td>
<td>7.3%</td>
</tr>
<tr>
<td>Intervals not covering their parameter: Out of the Selected</td>
<td>48%</td>
</tr>
</tbody>
</table>

Reporting the Pfizer-BioNTech COVID-19 Vaccine trial

**Note:** Number of participants at risk for the endpoint.

Credible interval for VE was calculated using a beta-binomial model with prior beta $(0.700102, 1)$.
Thresholds are helpful when actions are required. Comparing $P$-values to a significance level can be useful, though $P$-values themselves provide valuable information. $P$-values and statistical significance should be evaluated just as any other output of the statistical analysis.

Threshold for decision – action – or for selection

(Israel4 Cloud seeding experiment)

and selection is essential in modern science

Likelihood ratio, posterior odds, ..., are all practically subject to selection at a (sometimes) arbitrary threshold
'Scientists rise up against statistical significance'

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Amrhein, Greenland & McShane (’19)
They object to ‘Statistical Significance’ and eventually to any bright line

Relying on The Am. Statistician & Hurlbert, Lavine & Utts therein:

_Coup de Grâce for a Tough Old Bull:_

“Statistically Significant” Expires

They ‘ask’: “how can we address multiple comparisons without a threshold?” And answer: “We can’t. And should not try”.

Recommending instead:

“nuanced reporting” & “no need for bright line” as in Reifel et al ‘07
Reifel et al ‘17

Influence of river inflows on plankton distribution
Around the southern perimeter of the Salton Sea, California

Only results with $p \leq 0.1$

Are specifically discussed in the Abstract

Out of 41 results

Ban the use of Abstracts!
Capturing the uncertainty associated with statistical summaries is critical. Different measures of uncertainty can complement one another; no single measure serves all purposes. The sources of variation that the summaries address should be described in scientific articles and reports. Where possible, those sources of variation that have not been addressed should also be identified.
Addressing the relevant variability
Mouse phenotyping example: opposite single lab results

Figure 1 | Genotype-by-Laboratory interaction (GxL). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 (Supplementary Table 1). Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). a. illustrates significant genotype effect according to the Random Lab Model (RLM) with similar slopes indicating a small GxL effect. b. illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. c. exhibits substantial GxL: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such “opposite significant” (Supplementary Methods S1.1.3) cases were not rare using the standard method, but disappeared after GxL-adjustment. d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets: average single-lab Type-I error rate using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, see Supplementary Table 1.
GxL interaction is “a fact of life”

Genotype-by-Lab effect for a genotype in a new lab is not known; **but**

If its variability $\sigma^2_{GxL}$ can be estimated, use

$$
\frac{\text{Mean}(M_{G1}) - \text{Mean}(M_{G2})}{(\sigma^2_{\text{Within}} (1/n + 1/n) + 2\sigma^2_{GxL})^{1/2}}.
$$

We call it GxL- adjustment

It’s the relevant “yardstick” against which genetic differences should be compared, when concerned with replicability across laboratories.
**Addressing the relevant variability**

Mouse phenotyping example: opposite single lab results

![Graphs showing genotype-by-laboratory interaction](image)

**Figure 1 | Genotype-by-Laboratory interaction (G×L).** Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 (Supplementary Table 1). Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). a. Illustrates significant genotype effect according to the Random Lab Model (RLM) with similar slopes indicating a small GxL effect. b. Illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. c. Exhibits substantial GxL: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such “opposite significant” (Supplementary Methods S1.1.3) cases were not rare using the standard method, but disappeared after GxL-adjustment. d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets: average single-lab Type-I error rate, using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, see Supplementary Table 1.
Single-lab analyses in all known replication studies

6 Laboratories

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Open Field Test Path in Corners (cm)</th>
<th>Novel Object Test Exploration Time (min)</th>
<th>Elevated O Maze: Total Path (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BL/6</td>
<td>800</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>DBA/2</td>
<td>600</td>
<td>6</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p_{RLM} ≤ 0.002</th>
<th>p_{RLM} ≤ 0.01</th>
<th>RLM-NS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BL/6</td>
<td>DBA/2</td>
<td></td>
</tr>
</tbody>
</table>

2 Genotypes

Significant by: - G×L-adjustment - standard method - not significant

8 Multi-lab Datasets

<table>
<thead>
<tr>
<th>In-lab Significance of Non-replicable Results (Type I Error Rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td># Labs</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Figure 1 | Genotype-by-Laboratory interaction (G×L). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 (Supplementary Table 1). Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after G×L-adjustment (all at 0.05). a. illustrates significant genotype effect according to the Random Lab Model (RLM), with similar slopes indicating a small G×L effect. b. illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. c. exhibits substantial G×L: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such “opposite significant” (Supplementary Methods S1.1.3) cases were not rare using the standard method, but disappeared after G×L-adjustment. d. G×L-adjustment decreases non-replicable discoveries in 8 multi-lab datasets: average single-lab Type-I error rate using the standard t-test is much higher than the prescribed 5%. The G×L-adjustment brings it close to 5%, see Supplementary Table 1.
Utilizing large database to get $\sigma_{G\times L}$

Scientists conducting experiments in their lab can get an estimate of the relevant GxL variability from public Mouse Phenotyping Database e.g. Jackson Labs JAX, Bar Harbor)

“Replicability Adjuster” Implemented there

In current 3-labs experiment in TAU and Jax the proportion of non-replicable results reduced from 60% to 12%
In summary, $P$-values and significance tests, when properly applied and interpreted, increase the rigor of the conclusions drawn from data. Analyzing data and summarizing results are often more complex than is sometimes popularly conveyed. Although all scientific methods have limitations, the proper application of statistical methods is essential for interpreting the results of data analyses and enhancing the replicability of scientific results.

"The most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves, who keeps in the background the part he has played, perhaps unconsciously, in selecting and grouping them." (Alfred Marshall, 1885)
Thanks!

www.replicability.tau.ac.il

The industrialization of the scientific process